MITIGATING THE ERROR RATE OF AN IPF-BASED POPULATION SYNTHESIS APPROACH BY INCORPORATING MORE HETEROGENEITY INTO THE INITIAL SEED

Ismaïl Saadi
University of Liège
Department of Architecture, Geology, Environment and Construction (ArGEnCo)
Local Environment Management & Analysis (LEMA)
Quartier Polytech 1, Allée de la Découverte 9, BE-4000 Liège, Belgium
Tel.: +32 4 366 96 44; Fax: +32 4 366 29 09; Email: Ismail.Saadi@ulg.ac.be

Ahmed Mustafa
University of Liège
Department of Architecture, Geology, Environment and Construction (ArGEnCo)
Local Environment Management & Analysis (LEMA)
Quartier Polytech 1, Allée de la Découverte 9, BE-4000 Liège, Belgium
Tel.: +32 4 366 93 94 ; Fax: +32 4 366 29 09 ; Email: A.Mustafa@ulg.ac.be

Jacques Teller
University of Liège
Department of Architecture, Geology, Environment and Construction (ArGEnCo)
Local Environment Management & Analysis (LEMA)
Quartier Polytech 1, Allée de la Découverte 9, BE-4000 Liège, Belgium
Tel.: +32 4 366 94 99; Fax: +32 4 366 29 09; Email: Jacques.Teller@ulg.ac.be

Mario Cools, Corresponding Author
University of Liège
Department of Architecture, Geology, Environment and Construction (ArGEnCo)
Local Environment Management & Analysis (LEMA)
Quartier Polytech 1, Allée de la Découverte 9, BE-4000 Liège, Belgium
Tel.: +32 4 366 48 13; Fax: +32 4 366 29 09; Email: Mario.Cools@ulg.ac.be

Word count: 6,528 words text + 5 tables/figures x 250 words = 7,778 words

Submission date: November 15, 2016

ABSTRACT

In this paper, we propose a bi-level procedure for population synthesis to obtain good estimates of both the marginal and joint distributions. As a fitting-based algorithm, the Iterative Proportional Fitting (IPF) approach is capable of providing accurate synthetic population estimates from the marginal distributions perspective. The algorithm iteratively recalculates the different weights of the k-way contingency table until the deviation between the simulated and observed marginal distributions is minimized. Besides, the initial boundaries, referred to as the seed, are preserved in terms of proportions with respect to the full sample. This means that although the different values are updated along with the iterations, the correlation structure of the underlying population remains unaffected. In this paper, a hybrid model is proposed. Its originality resides in the incorporation of heterogeneity into the initial seed by using a Hidden Markov Model (HMM) *(1)*. The enriched initial seed is then fitted to a set of stable marginal distributions. This new hybrid approach is very interesting as it for instance can be calibrated by an unlimited number of PUMS, and separate information can be merged. The estimation of the correlation structure (synthetic seed) is improved compared to the standard seed (stemming from a micro-sample) thanks to the HMM. In this regard, the results show that the coupling of IPF with the HMM method provides better estimates, i.e. decreased RMSE of 79.16% for 1% sample size, from both the marginal and the joint distributions points of view. This implies that enriching the micro-sample is absolutely necessary before fitting with any aggregate marginal distributions.

*Keywords***:** Iterative Proportional Fitting (IPF), Hidden Markov Model (HMM)-based approach, Hybrid Model, population synthesis, agent-based micro-simulation modeling

1. BACKGROUND

The current need in extensively disaggregated data for agent-based micro-simulation modeling is important. The behavior realism in the simulation of urban and transportation systems depends deeply on the quality of the synthetic population. The major problem is that such disaggregated data are not always available. In addition, their sizes are generally limited (less than 10% in the best case for some countries). Travel surveys with detailed information about agents are costly and depending on the level of precision constrained by confidentiality. In this regard, aggregate data are used in addition to micro-samples of the full population.

Typically, three categories of population synthesis procedures can be distinguished in the literature: synthetic reconstruction, combinatorial optimization and generation-based methods.

The first category includes Iterative Proportional Fitting (IPF)-based approaches or methods derived from IPF, which are until now the main approaches used for modeling transportation and urban systems *(2–4)*. Basically, an IPF procedure consists of fitting a multi-dimensional contingency table given a set of marginal distributions. In practice, the marginal distributions are relatively stable and reliable. In this regard, they are used as controls such that the deviation between the simulated marginal distributions and the target ones are minimized. Generally the contingency tables are initiated by micro-surveys which correspond to sizes smaller than 10 % of the total population. Once the IPF algorithm converges, the corresponding contingency/multi-dimensional array can be used to generate a set of agents, i.e. individuals or households, using a Monte-Carlo simulation. As a fitting-based approach, IPF is particularly powerful in providing highly accurate synthetic populations, when match between the synthetic and observed populations are compared on the basis of the marginal distributions. In contrast, when the joint distribution of the simulated population is compared against the reference dataset, recent methods *(1, 5–7)* seem to be more appropriate.

Besides, to address the needs for small area micro-data in Britain, Voas and Williamson *(8)* proposed an improved reweighting approach, referred to as Combinatorial Optimization (CO). In essence, Combinatorial Optimization can be considered as an iterative procedure, where a combination of households is extracted from Samples of Anonymized Record (SAR) of households *(8)*, which represents 1% of the full population. These combinations are fitted against the marginal distributions, i.e. known small-area constraints, until the best fit is reached. The CO approach is a method adapted for avoiding problems related to the lack of geographical information inherent to microdata.

In a comparative study, Ryan et al. *(9)* outlined that both the IPF and CO approaches are capable of synthesizing small populations of firms. They also show that the accuracy of the synthetic populations increases with the level of tabulation details. Similarly, when the input sample size is important, the results tend to be better. In this regard, these trends are consistent with the findings of Saadi et al. *(1, 10)*. In addition, by using a similar amount of input data, the comparison reveal that CO-based approach provides better results than IPF. However, Ryan et al. *(9)* acknowledge that the comparison between CO and IPF includes some limitations such as the size of tested populations, i.e. tens of thousands, and the low level of scalability. Regarding the scalability, it is difficult to ensure that CO will show more stability against a significant increase in the number of cells. While the case study is relatively reasonable in terms of complexity, the computer run-time took around 12 to 15 mins for both methods. In this regard, dealing with higher population sizes might be problematic.

A third category of population synthesis concerns the generation-based approaches. In literature, three studies can be classified within this category: the simulation-based approach *(5)*, the Bayesian networks *(6)* and the Hidden Markov Models (HMM) *(1)*. Farooq et al. *(5)* proposed a methodology where partial views of the true joint distribution, in the form of conditional probabilities, are used to build a synthetic distribution by means of the Monte

Carlo Markov Chain (MCMC) algorithm. The conditional distributions are calibrated using different micro-samples. The advantage of the simulation-based approach resides in its ability to provide good estimates, despite the fact that only partial conditional distributions are used. A comparative study between IPF in the sense of Beckman et al. *(3)* and the simulation-based approach revealed that the latter method behaves better than IPF *(5)* and Iterative Proportional Updating (IPU) *(10)* for specific parameters settings, i.e. four attributes-based population synthesis.

Sun and Erath *(6)* used Bayesian Networks (BN) to represent the inter-dependencies in-between variables. In this regard, a BN can be assimilated to a graphical representation of the multivariate joint distribution of the considered population. The best model is selected based on AIC/BIC criterions. The BN approach is capable of capturing the configuration structure of the population efficiently *(6)*. A BN is quite stable with respect to scalability, curse of dimensionality, overfitting and resolution thanks to the estimation of local conditional distributions and directed acyclic graphs (DAG).

A recent approach based on HMM has been introduced by Saadi et al. *(1)*. This approach can be considered as a hybrid between the BN and MCMC approach. Indeed, Markov Chain principles are used in parallel with a strategy based on conditional distributions. In this regard, a HMM is a graphical representation of the configuration structure in the image of a BN. The advantages of BN and MCMC are merged within the HMM-based approach. In this regard, it offers a solid framework, providing accurate synthetic population estimates, while maintaining a limited level of data dependency. The approach of Saadi et al. *(1)* is characterized by its flexibility and efficiency in terms of data preparation. Information provided by an unlimited number of micro-samples (disaggregated data) can be captured by the HMM. Only one marginal distribution (aggregate data) is necessary to ensure a correct characterization of the true population. By using HMMs, Saadi et al. *(1)* revealed that for sample sizes lower than 25%, HMM outperforms IPF, even if the amount of input data is larger for IPF. In addition, the study of Farooq et al. *(5)* resulted in the same conclusion according to a comparison between IPF *(3)* and the MCMC approach. Analogously, the same conclusion is made by Sun and Erath *(6)* with respect to BN.

Besides, a wide range of more or less sophisticated approaches have been introduced in the literature. Some of them are mainly focusing on the matching between households and individuals using, e.g. a bi-partite graph *(11)*. In a comparative study, Lenormand and Deffuant *(12)* revealed that the sample-free approach established by Cargiulo et al. *(13)* provides better estimates than the sample-based approach of Ye et al. *(14)*. This can be explained by the fact that sample-based approaches, e.g. IPF, are particularly dependent on the initial seed. The dependency to the initial seed will be investigated in this paper.

In summary, fitting approaches such as IPF or IPU need aggregate data which are considered as reliable and quite stable inputs. In addition, one microdata sample is dedicated to initialize the multi-dimensional array. Without such initial microdata sample, an infinite number of combinations could fit the constraints which the array is subjected to. The proportions throughout the multi-dimensional are preserved although the values of cells are updated to mitigate the deviations between the simulated and observed targets. Fitting-based approach depends largely on the initial micro-sample which has been used to set the seed. As mentioned by Barthelemy et al. *(7)*, the correlation structure is preserved by IPF thanks to the odd ratios strategy. More details about the preservation of the weights within contingency tables can be found in Mosteller *(15)*. In this context, the values of the cells are updated until the target marginal distributions are fitted without running the risk of losing the configuration structure of the multi-dimensional array. In a way, this property of IPF can be considered as an advantage. In this regard, it is necessary to make the strong assumption of representability

of the micro-sample used as the seed by IPF. This assumption presents the risk that the synthetic population could strongly deviate from the true one.

To address this issue, we propose a hybrid model where the HMM-based approach *(1)* is coupled with the IPF algorithm in order to provide more accurate estimates for the population synthesis. Indeed, as mentioned previously, the HMM-based approach is capable of synthesizing a population with a lower error rate compared to IPF. However, as a fitting-based approach, IPF is better in providing a synthetic population with quasi-perfect marginal distributions unlike HMM. In contrast, the joint distribution is estimated less accurately. The hybrid approach has as objective to combine the strengths of IPF and HMM such that the synthetic population is improved from a global perspective, i.e. better marginal distributions while preserving accurate joint distributions.

The remainder of the paper is structured as follows. First, we describe the new methodology. Then, description of the dataset and corresponding data preparation steps are provided. Consequently, the results are discussed, and finally, the main conclusions with respect to the new methodology are formulated, and future research directions indicated.

2. METHODOLOGY

The lack of flexibility inherent to IPF and its dependency on the initial seed, which is estimated by a micro-sample, requires a more sophisticated approach such that data synthesis is made possible through different sources of information. In this way, the dependency to a single micro-sample can be avoided. In this paper, we propose a unique framework, where first a synthetic seed is generated using the HMM-based approach *(1)*, which is then used as an input for the IPF procedure. In doing so, the advantages of the HMM approach are combined with those of IPF. Figure 1 presents the different steps that should be followed to set up the methodology. The first step consists of collecting all the micro-samples (disaggregated data) and the marginal distributions (aggregated data) that are needed to calibrate the hybrid model. Then, depending on the case study, we need to identify which are the variables of interest. After that, the variables are classified according to their descending number of categories. Note that at this stage, the main settings that are being carried out are localized at the step dedicated to data fusion (Figure 1).

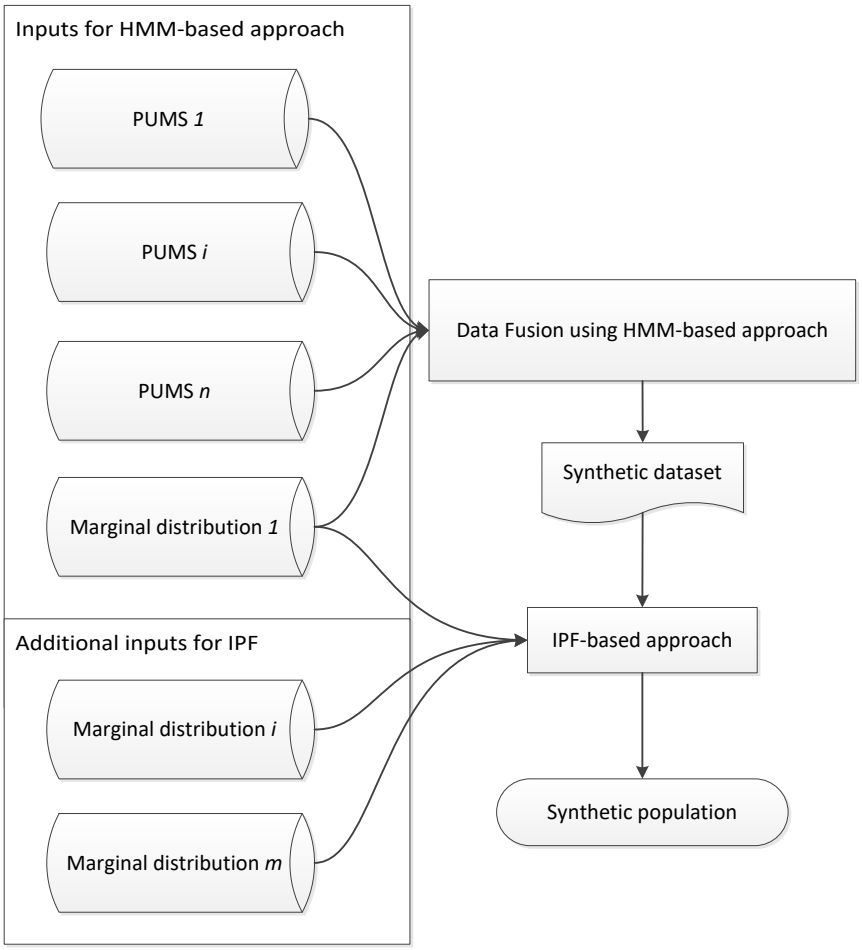

FIGURE 1 Representation of the overall framework

With respect to data fusion, all the transition probabilities are estimated in-between the different variables. In some cases, some readjustments of the categories, within each variable, need to be done to ensure sufficient consistency. When variables are aggregated, they should contain the same number of levels. Besides, it may happen that some variables are included in a microdata sample and not in another one. In this regard, the transition probability matrix can be estimated through different microdata samples. In contrast, when several transition probabilities associated to a specific combination of variables are available in more than one microdata sample, one could take the average values of the transition probabilities. A detailed procedure for dealing with multiple data sources with different samples sizes and incomplete dataset can be found in Saadi et al. *(1)*.

When the parameter settings of the HMM-based approach are estimated, a fixed number of agents attributes can be generated depending on the case study. In this context, a new synthetic dataset is obtained. This dataset is supposed to be a representative sample of the true population, which size can be set based on the size of the population to be synthesized. Figure 1 clearly shows that only one marginal distribution is needed, whereas all the microdata samples, e.g. PUMS, can be merged and included within the HMM framework. This means that valuable information with respect to other marginal distribution is ignored by the HMM. The coupling with IPF is a way to fill this gap, such that the information of multiple marginal distributions is taken into account.

Once the synthetic dataset is ready, the multi-dimensional array is estimated. After fitting the k-dimension contingency table, one could depict that the synthetic population

produced by the hybrid model is better than using both methods separately. This is illustrated by the case study.

As outlined by Saadi et al. *(1)*, the microdata samples can differ in terms of size and variable availability, illustrating the flexibility of the HMM-based approach. Note that one of the marginal distributions can be used twice within the framework. The HMM-based method uses a single marginal distribution, whereas IPF can use all the available marginal distributions. Alternatively, a partial fitting can be carried out as an alternative. With respect to the seed, i.e. the synthetic dataset, one should know that the HMM-based approach is capable of incorporating more heterogeneity. The problem of underestimation or absence of combinations of attributes with a low occurrence in the full population is solved by taking into account multiple data sources for estimating the initial seed such that a full heterogeneity is preserved.

In this paper, only the main features of the HMM-based synthetic population procedure are described to allow a replication of the above presented hybrid approach. In case of further details, one could refer to the work of Saadi et al. *(1)*. A HMM is a probabilistic framework based on a graphical representation of the links between attributes of a population. The idea behind an HMM-based approach consists of capturing and characterizing all these connections based on different data sources, e.g. surveys, where only a single source of aggregate information, i.e. a marginal distribution stemming from a census or statistics at national level, is necessary to generate a population.

When the variables of interest are selected, they should be classified according to their descending number of categories. For example, if we synthesize variables A, B and C with respectively two, ten and five categories, the formatted dataset should be established according to the following order: B, C then A.

Within the HMM-framework, two matrices need to be estimated: the transition probability *T* and the emission probability *E* matrices.

$$T = \begin{pmatrix} t_{11} & \dots & t_{1n} \\ \dots & t_{ij} & \dots \\ t_{n1} & \dots & t_{nn} \end{pmatrix}$$

where n is the total number of states of an HMM-based graphic and t_{ij} the probability to shift from states i to j. In the context of synthetic populations, this number corresponds to the sum of the categories of each attribute included in the synthesis procedure. For example, if one wants to synthesize two attributes, with respectively four and six categories, n will be equal to ten. In this regard, the transition probability matrix will be a 10×10 square matrix. A transition probability matrix can be estimated for each available microdata sample. At the end, a data fusion process can be done by incorporating the missing probabilities of a matrix into another one or by averaging some probabilities to reduce the variance.

Mathematically, the transition probabilities t_{ij} are estimated by using the following formula *(1)*:

$$t_{ij} = \frac{p_{ij}}{\sum_{k=1}^{N} p_{ik}}$$

where p_{ij} is the number of transitions shifting from state i, i.e. $S_{1,i}$ of variable A, to state j, $S_{1,j+1}$ of variable B. In the HMM-based framework, a state is defined as a category of a variable. p_{ik} is the number of transitions shifting from state i, i.e. $S_{1,i}$ of variable A, to all the

categories of the next variable, i.e. $S_{1,i+1} \rightarrow S_{m,i+1}$ with m the number of categories within variable B).

The emission probabilities play a particular role within each state. Depending on the nature of the variable (discrete or continuous) and the level of disaggregation, a probability emission can emit one or several symbols according to specific distribution. For example, gender has two states: male and female. In this particular case, the probability emission of the state "male" is 1. In contrast, if a state represents an interval, i.e. age between 18 and 35 years old, one could set the emission probabilities such that it follows a distribution calibrated on a given dataset including 18 and 35 as boundaries.

$$E = \begin{pmatrix} e_{11} & \dots & e_{1m} \\ \dots & e_{ij} & \dots \\ e_{n1} & \dots & e_{nm} \end{pmatrix}$$

where m is the value corresponding to the highest categorical variable and e_{ij} the probability that state i emits symbol j. In the above example, m is equal to ten. If one wants a fully disaggregated population synthesis such that all the categories remain the same, the emission probability should be concentrated into one cell and equal to one. Otherwise, one can determine a distribution depending on the nature of the application.

So far, we have shown how to integrate the disaggregated information coming from different microdata samples. At this point, we need to include the aggregate information, stemming from the marginal distribution, within the transition and emission probability matrices for consistency *(1)*.

When the parameters of the HMM-based approach are estimated, one can easily generate a fixed number of observations that fits the true population in terms of marginal and joint distributions. Note that the HMM-based method is a generation-based approach, contrary to IPF, which is a fitting-based approach.

The final step consists in ensuring that the simulated population is correctly estimated. Different metrics can be used to analyze the results from different perspectives. The R-square values, the slope and the SRMSE (Standardized Root Mean Square Error) can be used to determine the quality of synthetic populations *(1, 5, 6)*. Other metrics such as the RMSE (Root Mean Square Error) have also been used in various studies *(10, 16, 17)*. When comparing the simulated population to the observed one, at least two or three statistical metrics need to be calculated and put into perspectives to confirm the results. In this study, we will systematically compare the populations on the basis of the RMSE from both the marginal and the joint distributions perspective. In addition, the results of the HMM-based approach are also examined by means R-square values and slopes.

3. CASE STUDY

The dataset used in this study stems from the workforce survey of 2013 performed in Belgium. A set of 6 variables has been selected from the survey to calibrate the IPF, HMM and hybrid models. A spatial variable, which contains the highest number of categories (589 municipalities), four socio-demographics (age, gender, profession, status) and one transport-related variable (travelled distances) are included within the selection. In order to underline the influence of the sampling rate on the deviation between the simulated and observed populations, we took into account three sampling rates with respect to the full dataset (Table 1).

We suppose that the full dataset represents the full population of Belgium. In this regard, we can easily obtain the six marginal distributions that will all be used as aggregate information by IPF and partially used by the HMM model. Note that information about the three first categories of age is not included within the dataset. Indeed, young people and students are not surveyed with the workforce survey. In order to compare the evolution of the different proportions, Table 1 presents also the proportions of each category of each attribute according to three sampling rates, i.e. 1%, 5% and 10%. Generally such sampling rates are more pertinent, as it is quasi impossible in practice to get access to large-scale surveys, e.g. census, or micro-surveys, with a sampling rate higher than 10%.

When the sampling rate decreases, one could observe that some categories are not attributed. Indeed, the sampling rate is so low that the micro-sample does not cover all the combination of attributes. This problem is not related to the synthetic population procedure but to the random sampling procedure. In this case study, we will reveal how the hybrid model can incorporate more heterogeneity into the microdata sample, so that more combination of attributes can be captured.

Regarding the spatial distribution of the observations, the microdata sample of 1% covers only 33.28% of the total number of municipalities. The 5% and 10% microdata samples cover respectively 76.06% and 87.61% of the total number of municipalities. However, one should pay attention about the sampling rate of 1%. It is important to keep in mind that we are preserving a high level of disaggregation regarding the spatial attribute which contains 547 zones. The number of observations of the full dataset is 30,700, thus only 307 observations are included within the microdata sample of 1%. It means that it is not possible to cover more than 307 municipalities. In this context, we have also investigated higher sampling rates to mitigate the effect due to the limited spatial coverage of the zones.

In addition, the number of categories within the selected variables is also important. We notice that for discrete variables with a limited number of categories, the proportions are quite stable throughout the different sampling rates. However, problems related to missing values appear when the number of categories is important. For example, four categories (4-17-18-19) are not represented for age with respect to the full population, in the case of a 1% microdata sample. A similar conclusion can be drawn for the spatial attribute.

1 **TABLE 1 Distribution of Attributes With Respect to Different Sampling Rates**

Variables	Full population	1% microdata sample	5% microdata sample	10% microdata sample
Age	%	%	%	%
1	NA	NA	NA	NA
2	NA	NA	NA	NA
3	NA	NA	NA	NA
4	0.11	NA	0.13	0.13
5	1.64	0.65	1.37	1.63
6	6.44	5.86	5.80	6.03
7	8.97	6.19	9.12	8.63
8	9.86	9.77	9.06	9.61
9	10.36	12.05	10.88	10.46
10	11.22	13.03	12.18	11.30
11	11.65	12.05	12.44	11.63
12	12.36	13.68	12.05	12.12
13	10.97	9.12	10.49	11.34
14	9.00	10.10	8.53	8.93
15	5.23	6.19	5.80	6.16
16	1.83	1.30	1.76	1.60
17	0.24	NA	0.33	0.29
18	0.09	NA	0.07	0.13
19	0.03	NA	NA	0.03
Gender	%	%	%	%
Male	53.96	53.09	55.24	54.33
Female	46.04	46.90	44.76	45.67
Profession	%	%	%	%
1	26.08	22.80	25.08	25.67
2	40.94	42.02	41.43	41.07
3	16.54	18.89	17.00	16.87
4	7.43	8.47	7.75	7.46
5	5.30	3.91	4.95	4.95
6	3.39	3.91	3.52	3.62
7	0.32	NA	0.26	0.36
Status	%	%	%	%
1	4.28	4.89	3.52	4.30
2	4.32	4.89	4.69	4.50
3	4.42	3.91	3.91	4.59
4	5.23	4.89	4.76	5.18
5	11.32	14.01	12.38	11.43
6	14.93	12.38	15.18	15.08
7	10.59	12.05	11.66	11.73
8	3.77	4.89	3.78	3.49
9	20.11	17.92	19.93	19.32
10	2.94	3.58	2.93	3.06
11	0.66	0.65	0.39	0.55
12	0.52	0.98	0.84	0.59
13	2.94	2.93	2.93	2.80
14	11.52	9.77	11.07	11.01
15	1.72	1.95	1.43	1.66

16	0.72	0.33	0.59	0.72
Distance	%	%	%	%
1	7.97	7.49	8.40	7.88
2	4.82	4.23	4.10	4.46
3	5.48	5.54	5.80	5.64
4	4.37	6.84	4.56	4.63
5	8.07	8.80	8.99	9.25
6	3.97	2.28	3.91	4.14
7	3.63	4.23	4.17	3.91
8	3.67	2.61	2.93	2.30
9	1.32	0.65	0.98	1.21
10	5.89	5.54	5.41	5.54
11	1.32	2.28	1.63	1.56
12	3.14	3.26	3.19	3.36
13	1.52	1.30	1.30	1.24
14	1.10	0.98	0.85	0.98
15	5.07	6.19	5.41	5.50
16	1.07	0.65	0.78	0.65
17	1.37	0.98	1.11	1.24
18	1.22	1.63	1.43	1.17
19	0.40	0.33	0.39	0.39
20	4.17	4.23	4.23	4.01
21	0.45	NA	0.52	0.49
22	1.15	0.65	1.11	1.07
23	0.80	NA	0.33	0.59
24	0.50	0.65	0.46	0.42
25	3.74	5.21	4.63	4.40
26	0.47	0.65	0.79	0.52
27	0.55	0.33	0.59	0.55
28	0.50	0.33	0.33	0.46
29	0.14	NA	0.20	0.23
30	3.18	3.58	3.58	3.32
31	18.93	18.57	17.92	18.21
Geographical location	547/589	196/589	448/589	516/589

The highly disaggregated nature of certain variables, e.g. locations, age, distance, have been deliberately preserved to show how a merging of IPF with HMM can mitigate the error that can be caused by the effects of large multi-dimensional arrays. Keeping such a level of disaggregation can have a positive influence on specific applications in transport and urban systems. Indeed, simplification of a synthetic population problem by aggregating some variables can lead to a loss of valuable information *(1, 5)*.

4. RESULTS

In this section, we present the results of the synthetic population using a standard IPF-based approach within the meaning of Beckman et al. *(3)*, along with a comparison against the hybrid model that includes a richer seed. We propose an analysis from two perspectives by investigating the differences in marginal distributions between the simulated and observed populations and the differences in joint distributions between the simulated and observed populations using the RMSE.

The RMSE has been used in different studies *(10, 16–18)* to assess the quality of synthetic populations. It is defined according to the following formula:

$$RMSE = \sqrt{E(\tilde{\theta} - \theta)^2} = \sqrt{\sum \frac{(y_{predicted} - y_{reference})^2}{N}}$$

where $\tilde{\theta}$ and θ are respectively the simulated and observed populations and N the number of observations. Note that the same metric is used to measure accuracy of the marginal and the joint distributions.

The results presented in Table 2 reveal clearly the importance of the initial seed when using an IPF-based approach. In contrast, if we use the HMM-based approach to incorporate more heterogeneity within the initial seed, we notice that a significant improvement can be realized with respect to the error rate. Indeed, based on the RMSE, the structural configuration has been improved by 59.72% compared to the standard IPF-based procedure for generating a synthetic population. The initial sample, which has been randomly selected, measures 10% compared to the full population.

TABLE 2 Comparison Between HMM, IPF and Hybrid Models in Terms of RMSE

	HMM	IPF	Hybrid	Changes in % for the Hybrid model (with respect to IPF)	Changes in % for the Hybrid model (with respect to HMM)
Sample Size = 10%					
Municipalities	0.0003	0.0002	0.0003	+36.48	+3.45
Distances	0.0033	0.0023	0.0028	+21.74	-15.15
Age	0.0030	0.0028	0.0025	-10.71	-16.67
Status	0.0044	0.0036	0.0035	-2.78	-20.45
Profession	0.0028	0.0083	0.0024	-71.08	-14.29
Gender	0.0041	0.0148	0.0008	-94.28	-79.33
Joint	0.0312	0.0782	0.0315	-59.72	+0.96
Sample Size = 5%					
Municipalities	0.00051	0.00046	0.00028	-39.02	-44.56
Distances	0.0051	0.0031	0.0019	-38.71	-62.75
Age	0.0063	0.0029	0.0053	+82.76	-15.87
Status	0.0052	0.0047	0.0034	-27.66	-34.62
Profession	0.0056	0.0102	0.0047	-53.92	-16.07
Gender	0.0039	0.0079	0.0047	-40.51	+20.51
Joint	0.0315	0.1078	0.0364	-66.23	+15.56
Sample Size = 1%					
Municipalities	0.0024	0.0024	0.0020	-16.67	-16.67
Distances	0.0113	0.0113	0.0116	+2.65	+2.65
Age	0.0114	0.0124	0.0116	-6.45	+1.75
Status	0.0143	0.0154	0.0168	+9.09	+17.48
Profession	0.0116	0.0316	0.0434	+37.34	+274.14
Gender	0.0031	0.0070	0.0185	+164.29	+496.77
Joint	0.0306	0.1742	0.0363	-79.16	+18.63

Regarding the RMSE metrics related to the marginal distributions, one could depict that some important changes occurred between the standard model and the hybrid model. Indeed, the RMSE increased by +36.48% for the spatial attribute, which contains the most important number of categories (547). The same trend can be observed regarding the travelled distances classification with an increase of +21.74%. In contrast, the rest of the RMSE metrics seem to be decreasing when shifting towards the hybrid model. Note that the estimation of profession and gender is significantly improved by respectively -71.08% and -94.28%. These different results show that improving a model does not mean that all the marginal distributions should be improved from the RMSE point of view. In certain circumstances, it is more interesting to allow some minor deviations within some attributes, especially the ones that contain an important number of categories. In this context, the whole joint distribution can significantly be improved based on the RMSE.

In the case of a sampling rate of 5%, we notice that the hybrid model is capable of mitigating the error with respect to five out of six attributes, when compared to the HMM or IPF approaches. The changes are quite significant, especially for the municipalities (-39.02% and -44.56%). In contrast, IPF outperforms the hybrid approach for age, whereas the HMM approach outperforms the hybrid approach for gender. The hybrid model is capable of increasing indistinctly the RMSE of the joint distribution with respect to HMM (+15.56%) and decreasing by 66.23% with respect of IPF, whereas most RMSE related to the marginal distributions are improved. In this regard, the hybrid succeeds in finding a good trade-off between all types of errors.

In the case of a sampling rate of 1%, the sample is so small that the heterogeneity within the initial seed is too poor. Even though the marginal distributions are used as aggregate information, we notice that IPF is not capable of providing good estimates after applying the fitting procedure. If heterogeneity is incorporated into the initial seed, we observe an improvement in the error rates when the hybrid model is compared to the IPF-based approach (-79.16%). However, it is better to keep the synthetic dataset produced by the HMM-based approach, as all the RMSEs outperform IPF as well as the hybrid model.

Figure 2 presents the comparison between the marginal distributions related to the simulated and observed populations for the sampling rate of 10% that has been established by the hybrid model. One could depict that the matching between the different marginal distributions is very accurate, despite the fact that six attributes have been synthesized, even when an important level of disaggregation is kept within some of them.

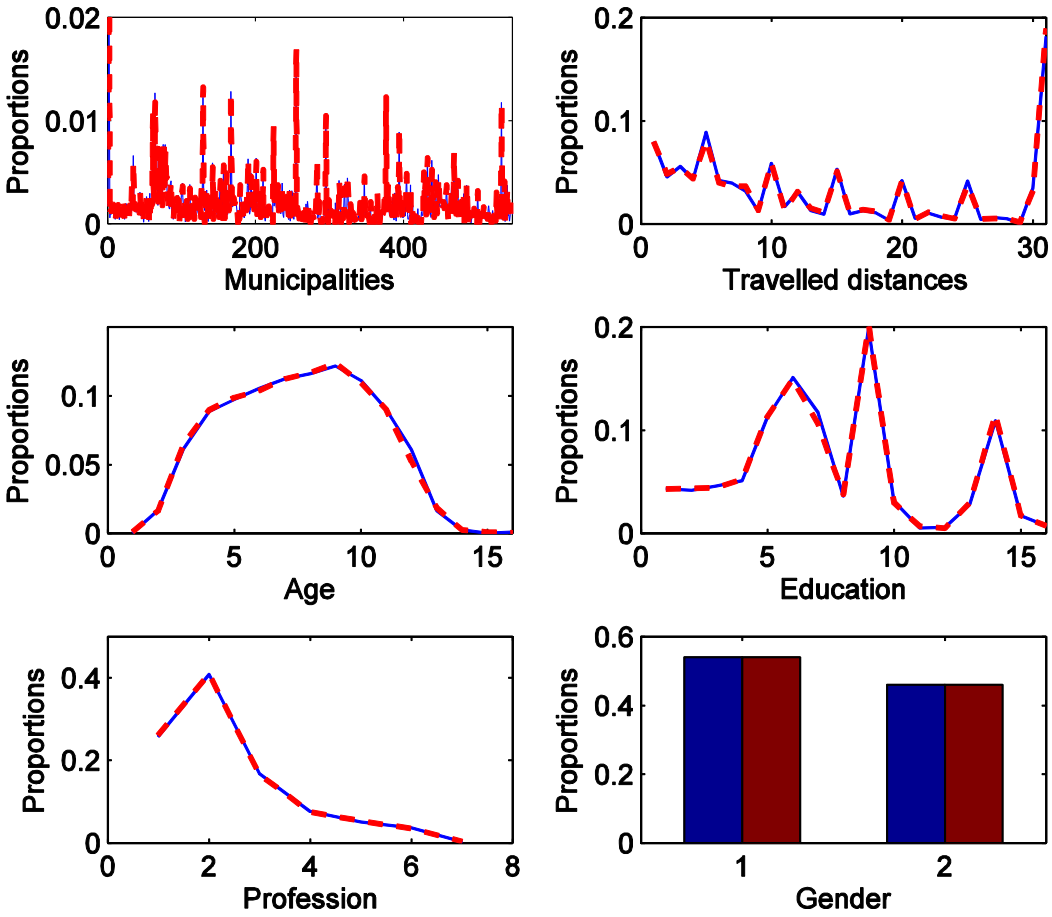

FIGURE 2 Comparison between the marginal distributions of the simulated (red broken line) and observed (blue continuous line) populations for the hybrid model (sample size=10%)

Figure 3 presents the connections in-between all the combinations of bi-variate joint distributions. In the current framework, we use the HMM-based approach to incorporate more heterogeneity into the initial seed. In this regard, we show an additional way for analyzing the quality of the graphical connections between the simulated and observed attributes. In this regard, it is strongly advised to examine the accuracy of these interconnections to ensure that the sub-model based on the HMM model is correctly calibrated. The R-square values and slopes are used to examine the quality of these inter-connections. We can observe that connections between attributes that are adjacently arranged provide better estimated in terms of R-square and slope. In contrast, the quality seems to decrease slowly for distant attributes.

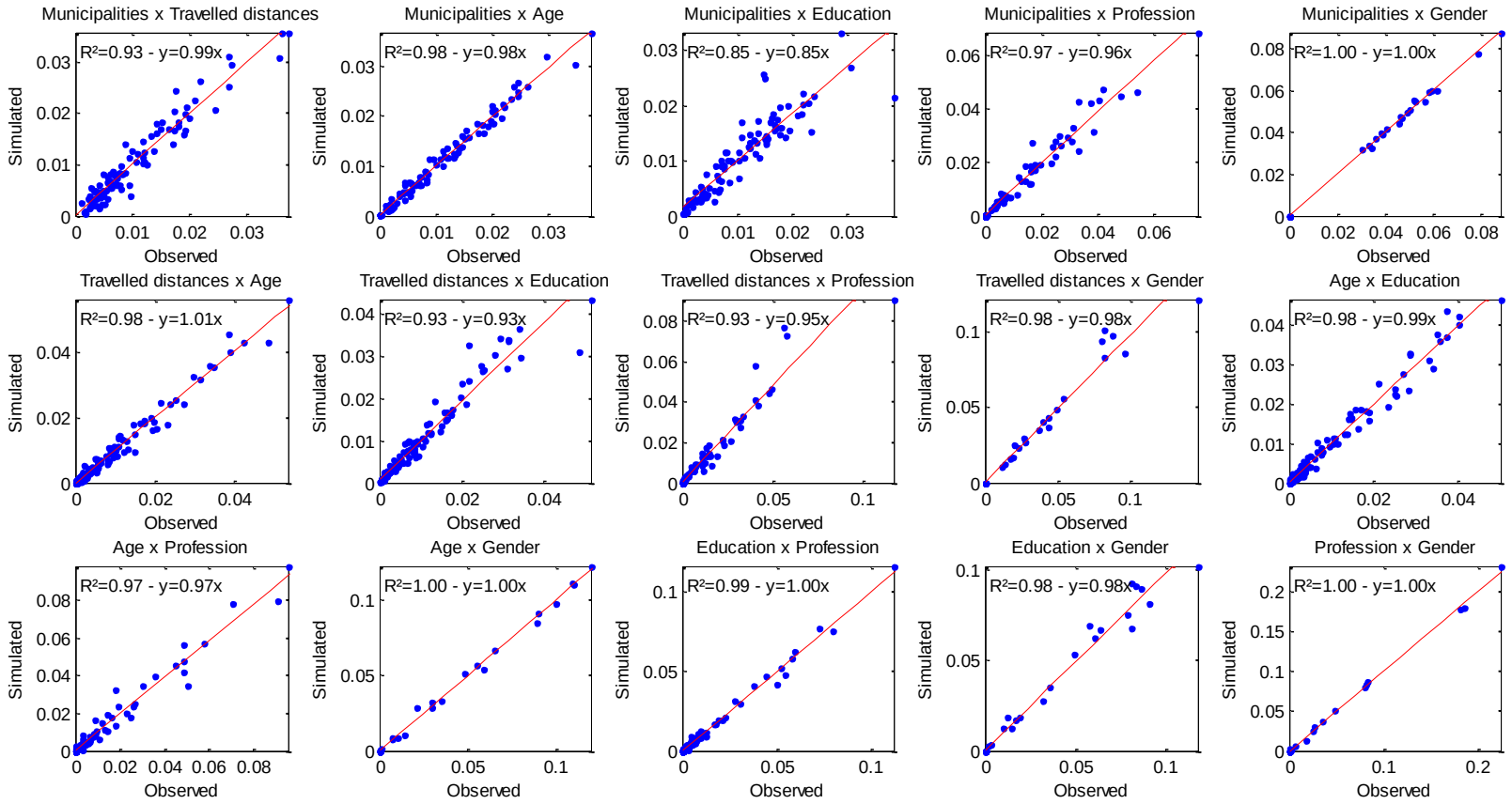

FIGURE 3 Comparison between the bi-variate joint distributions of the simulated and observed populations for the HMM-based model (sample size=10%)

Indeed, six variables have been synthesized: municipalities, distances, age, status, profession and gender based on the initial seed or PUMS. Following the recommendations of Saadi et al. *(1)*, the simulated and observed transition patterns are compared between each other by using R-square and slope values. These comparisons are very useful to assess the quality of the estimated transition patterns and make the detection of eventual anomalies within the population synthesis procedure easier.

By comparing the different sub-figures presented in Figure 3, one could depict that some of the fits show more dispersion, i.e. R-square=0.85 and y=0.85x, compared to other transition patterns where R-square values are very close or equal to 1. This phenomenon has been defined in Saadi et al. *(1)* as the dispersion generated by distances between variables. This is typically related to the graphical structure of the HMM-based approach. Indeed, in this framework, the variables are distributed according to the following order: municipalities (589 levels), distance (31 levels), age (19 levels), status (level of education) (16 levels), profession (7 levels) and gender (2 levels). By keeping in mind this configuration, the underlying dispersion can be easily understood. One could notice from the sub-figures that the bi-variate distributions formed by the adjacent variables present quasi-perfect R-square and slope values where Municipalities x distances (or travelled distances) is supposed to present the lowest R-square value, i.e. 0.93.

Besides, the figure related to Municipalities$\times$Education shows a low R-square compared to the rest of the results. Indeed, the drawing of the variable status depends only on the previous one, i.e. age with $P(\text{status}\,|\,\text{age})$. The same is true for age with distance $P(\text{age}\,|\,\text{distance})$, and also $P(\text{distance}\,|\,\text{municipalities})$. Instead of using full conditional probabilities, i.e. $P(\text{status}\,|\,\text{distance},\text{age},\text{municipalities})$, the structure of the HMM is established such that the conditional probabilities are systematically partial. Although the MCMC approach of Farooq et al. (2013) offers the possibility of using full probabilities, it presents an important dependency to the micro-sample (risk of overfitting) which is used for calibrating the conditional distributions (full or partials). In the HMM framework, dealing with conditional probabilities offer more flexibility and reduce the dependency to data. Moreover, this strategy allows for estimating the transition patterns from an unlimited number of datasets as explained in Saadi et al. *(1)*. The risk of overfitting is also avoided as the enriching of the initial seed will automatically be different from the original one. Particularly, the procedure of enrichment consists of generating observation sequences which do not exist in the original seed. The graphical structure of an HMM makes the incorporation of heterogeneity into the original microdata sample quite efficiently.

Globally, according to Figure 3, one could depict that for a sample of 10% (commonly used), the results related to the enriched microdata sample show that the transition patterns are accurately estimated. At this step, the synthetic microdata sample can be fitted to stable marginal distributions.

5. DISCUSSION AND CONCLUDING REMARKS

Different issues should be highlighted with respect to the different sub-modules, their interactions and the underlying results stemming from the case study. First, the current study confirms the findings of Saadi et al. *(1)*, who suggested that IPF outperforms the HMM-based approach in the context of a four attributes-based comparative study. We showed that the HMM-based approach performs better and shows more stability when the number of attributes that need to be simulated increases. This stability, that characterizes the HMM-based approach, is also shared within the hybrid model, as the error related to the multivariate joint distributions does not deviate significantly from the error rate determined for the HMM-based approach, although the hybrid model integrates the IPF procedure as a sub-module.

Regarding the complexity of the variables, in terms of the number of categories, we can see that the modeling framework can preserve the highly disaggregated nature of several variables with an acceptable error rate. A standard IPF cannot be sufficient to reach such a purpose as it would suffer strongly from the curse of dimensionality.

By merging IPF and HMM, all types of datasets (aggregate or disaggregate) can be included within the hybrid model. For example, if the HMM-based approach is calibrated separately, only a single marginal distribution (aggregate information) can be used, while additional aggregate information cannot be integrated from a technical point of view. In this regard, the hybrid model offers the possibility to take into account additional source of information at an aggregate level through the IPF-based module. Inversely, only one microdata sample can be integrated as the initial seed within IPF. We have explained in the section related to the methodology that information stemming from an unlimited number of microdata samples could be extracted to estimate step by step the transition probability matrix of the HMM-based approach *(1)*. In this way, the hybrid model succeeds in merging the information and providing synthetic populations, while maintaining a good trade-off between the RMSE of the marginal and multivariate joint distributions. Besides, the preparation of an accurate initial seed is fundamental and has a significant influence on the final synthetic population. Fitting quasi-perfectly the marginal distributions (aggregate data) should not be a goal in itself. Sometimes, a very good fit from the marginal distributions point-of-view may decrease the quality of the configuration structure of the synthetic population with respect to the true one. In this regard, IPF presents a bad capacity in adjusting the contingency table such that it is a good representation of the true population. Particular attention is necessary in building a correct and accurate initial seed.

Besides, one could conclude from the results that, in the case of IPF, the quality of the initial seed strongly affects the error rate of the multi-variate joint distributions, i.e. -59.72 % of deviation with respect to IPF when the hybrid model is taken as the reference. The deviation is still increasing when the sampling rate decreases. We can observe that the hybrid approach succeeds in finding a good equilibrium between a significant improvement of the RMSE (marginal distributions) compared to IPF, while keeping an accurate value of RMSE (joint distribution) compared to HMM as well.

As IPF is used as a sub-module, the zero-cell effect can be mitigated or avoided by adopting more sophisticated approaches *(19)*. In this paper, we want to focus on the importance of the initial seed with its underlying problematic issues, i.e. lack of heterogeneity, small initial sample size. In this regard, the bad results of IPF are also due to the important dependency to the initial micro-sample, as shown through the different case studies.

According to the results, it is clearly interesting to incorporate more heterogeneity in the initial sample by implementing, e.g. an HMM-based approach. Then the synthetic dataset produced can be used as a synthetic initial seed for the IPF-based approach to adjust the contingency table from the marginal distributions point of view. Note that the problem related to the size of the initial population does not exist anymore in the hybrid model, as the HMM-based is a generation-based technique. In this regard, a fixed number of observations can be generated from the HMM model such that the characteristics of the synthetic dataset are preserved.

The strategy that integrates simultaneously the HMM and the IPF-based approaches works very well. The results presented in Table 2 for the 5% and 10% sample sizes show that the hybrid model is capable of mitigating the error of the configuration structure with respect to standard IPF. In addition, the errors of the marginal distributions are, in general, also mitigated with minor exceptions. However, in the case of a 1% sample size, the trends are not really similar to those presented for higher sample sizes. This can be explained by the way the case study has been defined. Indeed, as we supposed that the workforce survey of 2013

represents the whole population of Belgium, a sampling rate of 1% is associated to 307 individuals. In such a situation, the problem is not related to the integrated approach itself but to the default in the sampling procedure. Indeed, the number of municipalities is higher than 307, i.e. 589. Thus, the sampling rate does not even cover all the municipalities. Given the fact that the spatial variable "municipalities" is used as the constraint by the HMM, some of the municipalities are not present in the learning dataset of 1% that need to be enriched. In this context, it is not possible for the HMM-based approach to establish the transition patterns between the missing municipalities and the rest of the variables although the aggregate variable "municipalities" is used as constraint. In brief, the incorporation of heterogeneity into the initial seed gives, once again, to the hybrid-based approach a better stability for any sample size unless default in sampling procedure are present (see 1% sampling rate situation). However, according to the findings of the current paper, it is particularly recommended to enrich the initial sample by using an HMM-based approach. Alternative approach could be adopted such as Bayesian Networks *(6)* or MCMC *(5)*. However, further studies need to be carried out to ensure that those approaches have the capacity to incorporate heterogeneity into poor microdata sample.

6. ACKNOWLEDGEMENT

The research was funded by the ARC grant for Concerted Research Actions for project no. 13/17-01 entitled "Land-use change and future flood risk: influence of micro-scale spatial patterns (FloodLand)" and by the Special Fund for Research for project no. 5128 entitled "Assessment of sampling variability and aggregation error in transport models", both financed by the French Community of Belgium (Wallonia-Brussels Federation).

7. REFERENCES

1. Saadi, I., A. Mustafa, J. Teller, B. Farooq, and M. Cools. Hidden Markov Model-based population synthesis. *Transportation Research Part B: Methodological*, Vol. 90, 2016, pp. 1–21.
2. Arentze, T., H. Timmermans, and F. Hofman. Creating Synthetic Household Populations: Problems and Approach. In *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2014, Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 85–91.
3. Beckman, R. J., K. A. Baggerly, and M. D. McKay. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, Vol. 30, No. 6, 1996, pp. 415–429.
4. Zhu, Y., and J. Ferreira. Synthetic Population Generation at Disaggregated Spatial Scales for Land Use and Transportation Microsimulation. In *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2429, Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 168–177.
5. Farooq, B., M. Bierlaire, R. Hurtubia, and G. Flötteröd. Simulation based population synthesis. *Transportation Research Part B: Methodological*, Vol. 58, 2013, pp. 243–263.
6. Sun, L., and A. Erath. A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, Vol. 61, 2015, pp. 49–62.
7. Barthelemy, J., and P. L. Toint. Synthetic Population Generation Without a Sample. *Transportation Science*, Vol. 47, No. 2, 2013, pp. 266–279.
8. Voas, D., and P. Williamson. An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *International Journal of Population Geography*, Vol. 6, No. 5, 2000, pp. 349–366.

9. Ryan, J., H. Maoh, and P. Kanaroglou. Population Synthesis: Comparing the Major Techniques Using a Small, Complete Population of Firms. *Geographical Analysis*, Vol. 41, No. 2, 2009, pp. 181–203.
10. Saadi, I., A. Mustafa, J. Teller, and M. Cools. Forecasting travel behavior using Markov Chains-based approaches. *Transportation Research Part C: Emerging Technologies*, Vol. 69, 2016, pp. 402–417.
11. Anderson, P., B. Farooq, D. Efthymiou, and M. Bierlaire. Associations Generation in Synthetic Population for Transportation Applications. In *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2429, Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 38–50.
12. Lenormand, M., and G. Deffuant. Generating a Synthetic Population of Individuals in Households: Sample-Free Vs Sample-Based Methods. *Journal of Artificial Societies and Social Simulation*, Vol. 16, No. 4, 2013, p. 12.
13. Gargiulo, F., S. Ternes, S. Huet, and G. Deffuant. An Iterative Approach for Generating Statistically Realistic Populations of Households. *PLoS ONE*, Vol. 5, No. 1, 2010, p. e8828.
14. Ye, X., K. C. Konduri, R. M. Pendyala, B. Sana, and P. Waddell. Methodology to Match Distributions of Both Household and Person Attributes in Generation of Synthetic Populations. In *Proceedings of the 88th Annual Meeting of the Transportation Research Board*, Transportation Research Board of the National Academies, Washington, D.C., 2009.
15. Mosteller, F. Association and Estimation in Contingency Tables. *Journal of the American Statistical Association*, Vol. 63, No. 321, 1968, pp. 1–28.
16. Saadi, I., A. Mustafa, J. Teller, and M. Cools. An Integrated Framework for Forecasting Travel Behavior Using Markov Chain Monte-Carlo Simulation and Profile Hidden Markov Models. In *Proceedings of the 95th Annual Meeting of the Transportation Research Board*, Transportation Research Board of the National Academies, Washington, D.C., 2016.
17. Vovsha, P., J. E. Hicks, B. M. Paul, V. Livshits, P. Maneva, and K. Jeon. New Features of Population Synthesis. *Proceedings of the 94th Annual Meeting of the Transportation Research Board*, Transportation Research Board of the National Academies, Washington, D.C., 2015.
18. Lee, D.-H., and Y. Fu. Cross-Entropy Optimization Model for Population Synthesis in Activity-Based Microsimulation Models. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2255, 2011, pp. 20–27.
19. Guo, J. Y., and C. R. Bhat. Population Synthesis for Microsimulating Travel Behavior. In *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2014, Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 92–101.